U.S. Army Research Institute
for the Behavioral and Social Sciences

Research Report 1898

Mental Models for Effective Training: Comparing Expert and Novice Maintainers' Mental Models

Robert C. Hubal
RTI International

May 2009

U.S. Army Research Institute
for the Behavioral and Social Sciences

A Directorate of the Department of the Army
Deputy chief of Staff, G1

Authorized and approved for distribution:

Michelle Sams

MICHELLE SAMS
Director

Research accomplished under contract
for the Department of the Army

Research Triangle Institute International

Technical review by

Marshell G. Cobb, U.S. Army Research Institute
Thomas R. Graves, U.S. Army Research Institute

NOTICES

REPORT DOCUMENTATION PAGE

1. REPORT DATE (dd-mm-yy) May 2009	2. REPORT TYPE Final	3. DATES COVERED (from. . . to) July 2008 - January 2009
4. TITLE AND SUBTITLE Mental Models for Effective Training: Comparing Expert and Novice Maintainers' Mental Models		5a. CONTRACT OR GRANT NUMBER W74V8H-04-D-0044 DO 0006
		5b. PROGRAM ELEMENT NUMBER 633007
6. AUTHOR(S) Robert C. Hubal (Research Triangle Institute International)		5c. PROJECT NUMBER A792
		5d. TASK NUMBER 360
		5e. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) RTI International 3040 Cornwallis Road, PO Box 12194 Research Triangle Park, NC 27709-2194		8. PERFORMING ORGANIZATION REPORT NUMBER RTI Project #0209586.003
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES) U.S. Army Research Institute for the Behavioral and Social Sciences 2511 Jefferson Davis Highway Alexandria, VA 22202-3926		10. MONITOR ACRONYM
		11. MONITOR REPORT NUMBER Research Report 1898

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution is unlimited.

13. SUPPLEMENTARY NOTES
Contracting Officer's Representative and Subject Matter POC: William R. Bickley

14. ABSTRACT *(Maximum 200 words)*:
For a well-defined domain of knowledge, the process of learning can be characterized as a student's construction of a mental model of the domain's elements and their inter-relationships. It follows that a student's mental model can be compared against a desired mental model (such as an expert's) before, during, and at the conclusion of instruction. Differences between mental models could be useful for student diagnosis and assessment and for curriculum modification. In this investigation mental models of test, measurement, and diagnostic equipment (TMDE) usage for novice, intermediate, and expert U.S. Army ordnance electronics maintenance personnel were characterized and compared. Comparisons revealed differences between non-experts and experts, and also differences among experts.

15. SUBJECT TERMS
mental models; test maintenance and diagnostic equipment (TMDE); ordnance electronics training; training assessment

SECURITY CLASSIFICATION OF			19. LIMITATION OF ABSTRACT	20. NUMBER OF PAGES	21. RESPONSIBLE PERSON
16. REPORT Unclassified	17. ABSTRACT Unclassified	18. THIS PAGE Unclassified	Unlimited	37	Diane Hadjiosif, Technical Publications Specialist 703-602-8047

Research Report 1898

Mental Models for Effective Training: Comparing Expert and Novice Maintainers' Mental Models

Robert C. Hubal
RTI International

ARI – Ft Benning Research Unit
Scott E. Graham, Chief

U.S. Army Research Institute for the Behavioral and Social Sciences
2511 Jefferson Davis Highway, Arlington, Virginia 22202-3926

May 2009

Army Project Number	**Personnel Performance and**
633007A792	**Training**

ACKNOWLEDGEMENT

The author wishes to thank Dr. Will Bickley of ARI for critical guidance and support throughout this effort. Ms. Margaret McDowell of RTI was instrumental in identifying the online tool used during the experiment, in refining stimulus items used during the experiment, and in conducting the experiment. Mr. Steve Luthringer of RTI served the key role of in-house subject-matter expert.

Grateful thanks are due to LTC Edward Buckner, USA, and Mr. Gerald Holshouser of the U.S. Army Ordnance Electronics Maintenance Directorate at Ft. Gordon, GA. Without their generous assistance and support in providing access to subject-matter experts, to instructors, to students in residence, and to training resources and materials, this work would not have been possible.

MENTAL MODELS FOR EFFECTIVE TRAINING: COMPARING EXPERT AND NOVICE MAINTAINERS' MENTAL MODELS

EXECUTIVE SUMMARY

Research Requirement:

For a well-defined domain of knowledge, the process of learning can be characterized as a student's construction of a model of the domain's elements and their inter-relationships. This mental model is a hypothesized structure which the student creates and then actively consults and modifies while interacting within the domain.

To the extent that characterization of such a model is reliable and efficient (for an instructor, who must both construct a mental modeling instrument and administer the instrument), it could be useful for monitoring student progress during training, for assessing student end-of-course outcomes, and, when viewed for common anomalies across students, for making indicated modifications to course curricula. There is, then, the potential for characterization of mental models to be a valuable assessment tool for institutional Army training of well-defined domains.

Procedure:

Novice, intermediate, and expert electronics maintenance Soldiers of the Ordnance Electronics Maintenance Training Department at Ft. Gordon participated in an open sort task. The task consisted of sorting 39 test, maintenance, and diagnostic equipment (TMDE) stimulus items (TMDE descriptions and functions) into Soldier-specified categories. Sorts within groups and across groups were subjected to both qualitative and quantitative analyses.

Findings:

A qualitative inspection of the sorting data indicated explicable differences between descriptive and functional items for novices, intermediates, and experts.

A quantitative multiple dimensional scaling of the sorting data yielded three dimensions of categorizing across the groups. Weightings along one of the dimensions showed that experts differed from novices for functional items but not descriptive items, that experts differed from intermediate participants for descriptive items but not functional items, and that intermediate participants differed from novice participants for both descriptive and functional items. Individual randomly selected novice and intermediate Soldiers were found to differ from experts.

Utilization and Dissemination of Findings:

These results will be used to direct future investigations of mental models as a diagnostic measure for institutional training. The results were briefed to the U.S. Army Ordnance Electronics Maintenance Directorate at Ft. Gordon, GA, in March 2009.

MENTAL MODELS FOR EFFECTIVE TRAINING: COMPARING EXPERT AND NOVICE MAINTAINERS' MENTAL MODELS

CONTENTS

Page

LIST OF TABLES

LIST OF FIGURES

Mental Models for Effective Training: Comparing Expert and Novice Maintainers' Mental Models

Introduction

Structure of Representations

This research addresses the general topic of assessment of learning, in particular, the assessment of selected basic Army electronics maintenance skills. Typically, learning is assessed via observation of behavior or performance subsequent to instruction: Can the student exhibit the behavior or performance that was targeted by instruction? A somewhat different approach to assessment is to attempt to characterize the mental representation the student may have formed that allows production of the targeted behavior or performance. That is, is the student's representation of the learning domain sufficient to allow him or her to exhibit successfully the desired behavior or performance?

In cognitive psychology, *representation* is a catch-all word; knowledge and skills get represented. Interesting research questions what knowledge or skills (such as objects, situations, tasks, and strategies) get represented, how they get represented, and how the representations are made useful.

A representation models or depicts or portrays or delineates, suggesting the variety of knowledge and skills to be represented. For cognitive psychologists, a representation makes knowledge and skills accessible and modifiable. Typically, if a researcher requires indirect techniques to access the structure of representations then they are internal to a given participant. Different indirect techniques, described next, that reveal the reasonable representation of internal knowledge and skills include sorting and categorization tasks, similarity scaling tasks, and memory and estimation or inference tasks (Chase & Simon, 1973; Chi, Feltovich, & Glaser, 1981; Chiesi, Spilich, & Voss, 1979; Cooke & Schvaneveldt, 1988; Freyhof, Gruber, & Ziegler, 1992; Murphy & Wright, 1984).

Sorting and categorization tasks require a participant to assign commonality among elements in a set of domain-related items, according to his or her own criteria. The resulting sets are usually analyzed using multidimensional scaling, resulting in a "space" where like items cluster along *dimensions*. Dimensions can sometimes be labeled with descriptive phrases to indicate how participants array the elements. For example, an investigator might expect participants to categorize common fruits along a color dimension, a shape dimension, and perhaps a citrus/non-citrus dimension. More to the point for the research presented here, *a priori* the investigator might have some reason to believe Soldiers with little electronics maintenance experience would cluster certain test, measurement, and diagnostic equipment (TMDE) together because of their structural similarity, while Soldiers with considerable experience would cluster different TMDE together because of their diagnostic or functional similarity. One or more relevant dimensions representing structural, diagnostic, and/or functional elements could then derive from the multidimensional analysis.

Similarity scaling tasks require participants to characterize the relationships among domain elements, often by judging pairwise similarity between elements. Clustering analyses of

these data yield a kind of picture, not unlike a semantic network, where clusters represent “like” items and links suggest the relationships between clusters (e.g., strongly related, weakly related, negatively related, unrelated). Using the resulting clusters, it is relatively easy to compare or contrast networks, such as between expert and novice participants. Measures such as the contents of clusters, focal clusters, the density of links, interconnectedness, and link strength differences readily derive from cluster and network analysis software.

Memory and estimation or inference tasks are used to understand functional relationships among domain elements. Typical memory tasks such as free recall and reconstruction allow the participant to portray the schemas according to which he/she encodes the elements, and provide concrete means (e.g., via accuracy measures) to compare performance between participants. The seminal example comes from Chase and Simon (1973) where expert chess players reconstructed board positions (and were generally successful) not piece by piece but in related groups, whereas novice chess players reconstructed board positions piece by piece (and were no more accurate than would be predicted by memory span limits). Another example comes from Vicente (1992), on a task in which participants viewed process-control simulations and estimated final process-control variable values, where expert mechanical engineering graduate students estimated variable values better than novice non-engineering graduate students when inputs to the system were meaningful, but not when random. These kinds of tasks inform the investigator of models that the participants might be following to produce the items recalled or the estimates made.

Differences in Representations

Mental representational structures can have implications for task performance, particularly when they differ between experts and novices. For instance, Hassebrock et al. (1993), using patient protocol sheets, found that, after a delay, novices recalled information in its original format, whereas experts recalled mainly diagnostic information. This finding mirrors Chase and Simon in that experts but not novices were able to derive additional meaning (“chunks” of pieces in the case of Chase & Simon, diagnostic information in the case of Hassebrock et al.) from what was originally presented. Cooke and Schvaneveldt (1988; see also Chi et al., 1981) produce similar findings for programming experts: expert mental representation differs from novice representation. Chi et al. claim that expert/novice differences in representation stem from poorly formed, qualitatively different, or missing category knowledge in novice participants. Hassebrock et al. claim that problem-solving success requires links, supposedly formed through experience, between problem information and existing strategies. The important finding overall is that experts and novices mentally organize knowledge and skills differently.

Not only do experts and novices differ, but also experts can differ from each other. Smith (1990) found that different types of experts organize domain knowledge differently (see also McGraw & Pinney, 1990; Weiser & Shertz 1983). Smith presented two types of biology experts, faculty and genetics counselors, as well as biology novices, with genetics problems to categorize and solve. Faculty and counselors solved the same number of problems (and more than novices), yet they categorized differently. Smith argued that experts represent knowledge to facilitate its use; when different expert types have different purposes, they employ different knowledge representations. Relatedly, Weiser and Shertz (1983) found that novice and expert programmers categorized programs differently, experts focusing on algorithmic features (more meaningful)

and novices focusing on application area (a more superficial feature). More interestingly, Weiser and Shertz found that managers focused instead in terms of who would write the programs, a pragmatic feature that is relevant for them but not for the other participants.

Mental Models

A typical mental model employs the knowledge structures derived from indirect techniques as part of an active process of trying to understand a system's performance. That is, an individual's internal representation of knowledge and skills is actively consulted while interacting within the domain of interest. For instance, Norman (1983, 1988) describes an action cycle to describe how individuals interact with complex systems. Given a goal, such as needing to perform a task on a system that is generally understood but not in full detail, Norman suggests that individuals form an intention to act, identify a sequence of actions, act, perceive what happened, interpret what they perceive, and evaluate the interpretation as it informs the goal. More specifically related to the current research, the individual would rely on his/her representation of the system to inform each stage. Thus at the intention to act stage, the individual needs information regarding what actions are allowed, what each action accomplishes, and what comprises each action. For a particular piece of TMDE, the novice Soldier's available test actions would likely be a small subset of the expert's, leading to lesser assessment of system capabilities and thus different perceptions of what the actions accomplish. Similarly, at the interpretation stage, when the individual verifies whether or not the results of an action are appropriate, the individual relies on his/her current knowledge of system outputs, limitations, and interdependencies. In the TMDE example, the expert Soldier would have a much better understanding of test results and implications than the novice.

There have been a number of studies of differences between mental models of experts and novices, and occasionally between different types of experts. For instance, Gott, Bennett, and Gillet (1986) analyzed how electronics technicians solved certain technical problems, by considering the technicians' goal structures, understanding of context and tasks involved, and initial problem representations. Gott et al. found that having more accurate models of how the system functioned, as well as what and how tasks needed to be solved, led to better performance. Similarly, Kieras and Bovair (1984) were interested in operation of a device's control panel and the understanding of how the controls affected the device's operation. Kieras and Bovair showed that first learning a model of the device's internal mechanisms led to better performance, compared to simply learning the steps of operation by rote. Also, Hanisch, Kramer, and Hulin (1991) regarded how understanding the relations among system features influences the understanding of how to control the system. Meanwhile, Hagemann and Scholderer (2007) demonstrated differing views between growers and consumers of benefits and risks associated with genetically-modified food ingredients. Finally, Barnard, Reiss, and Mazoyer (2006) studied participants' mental models of documentation associated with a system and how different types of instruction (including demonstrations) are perceived differently by experts and novices.

Given that there are differences between novices' and experts' mental models, it follows that an individual's mental model can change as a function of the individual's interactions within the domain of interest. Over the course of repeated interaction (or instruction), an individual would be expected to build and update his or her model based on the results of those interactions.

Usefulness of Representational Structures

A line of research into mental models considers individuals' representations of structural, behavioral or mechanistic, and functional aspects of a system (Goel et al., 1996; Hmelo-Silver & Pfeffer, 2004). Structural models would be generally context free, dealing with system components but not influenced by the nature of specific tasks associated with the use of the system. Behavioral and functional models connect tasks and actions and would be context dependent, particularly in the case of functional models that could serve as strategic bases for understanding how to apply or employ the system in a given context. Research often suggests that mental representations focus on only a subset of types of components of a system (e.g., structural and not behavioral or functional), and that novices often represent what is readily (perceptually) available and static while experts are able to integrate the different system components (see also Chi et al., 1981). Along these lines, Soldiers who don't fully understand TMDE may be focused on the appearances of test equipment (regardless of their function) or procedural skills (that is, overt behavioral aspects of the equipment) to the exclusion of (meaningful) functional aspects of the equipment that would help elaborate their mental models. A training program using mental models is quite feasible for a domain that has well-defined processes and relationships, where the modeling can be performed reliably and reasonably efficiently. The specific domain involved in the current effort, Soldiers' understanding of electronic test equipment functionality, fits those criteria. It is reasonable to conceive of mental models as (1) reliably characterizing a Soldier's understanding of processes and relationships inherent in a given domain, (2) changing over time as the Soldier is trained, and (3) valid enough to gauge a Soldier's understanding in comparison to an expert's. Given these conjectures, it makes sense that a training program could be devised using mental models as a means of assessment for remediation and forward recommendations. The selection of a technique or techniques to capture and compare expert and novice mental models is seen as the first step in the development of an automated or semi-automated instrument to assist instructors or instructional systems in deriving mental models. Ultimately the Army may be able to adapt training to enable tailored remediation for a Soldier or recommendations for subsequent course modules given a Soldier's performance.

Electronics Maintenance

In the current work the knowledge and skills involved were related to use of TMDE during maintenance training of electronics systems. The intent was to model the structure of knowledge and skills exhibited by electronics maintenance personnel and to identify differences between experts' models and novices' models. The structure of representations was measured using the indirect technique of stimulus item categorization. The potential usefulness of this work lies in the capability to exploit any deltas between expert and novice representations.

The modeling work was conducted with the U.S. Army Ordnance Electronics Maintenance Training Department (OEMTD) located at Ft. Gordon, GA. The department is aligned under the Director of Training U.S. Army Ordnance Munitions and Electronics Maintenance School (OMEMS) located at Redstone Arsenal, AL. OEMTD students are assigned to 73rd Ordnance Battalion, a training battalion collocated with the 15th Regimental Signal Brigade at Ft. Gordon. The battalion's higher headquarters is the 59th Ordnance Brigade, located at Redstone Arsenal.

The training for OEMTD's six military occupational specialties (MOS) 94D, 94E, 94F, 94L, 94R, and 948B (for Warrant Officers) is housed at Ft. Gordon. The throughput as of mid-2008 was some 800 Soldiers per year and climbing.

Training lasts between 18 and 32 weeks, depending on MOS. All Soldiers run through an initial computer aided instruction (CAI). Soldiers run through the lessons at their own pace (not as part of a cohort). This instruction is for basic electronics principles and involves both simulated and hands-on components, and has a number of checks on learning. The training uses off-the-shelf equipment as well as training-specific equipment such as a test panel for testing preset circuit boards. The CAI has been used by OEMTD for about three years and, according to OEMTD personnel, is undergoing some redesign to address concerns, such as Soldiers completing the training but not understanding the reasoning behind some basic electronic tests.

A Soldier who completes the CAI is ready for MOS-specific training. OEMTD estimates that 80% or so of the content to be learned by Soldiers during MOS-specific training is common across all MOS. Even so, once completed with CAI, Soldiers separate into different classes for MOS-specific training. This cohort then receives both the common training and specialized training needed for the MOS. The training is given in classrooms and is both lecture and hands-on.

At issue is the understanding that Soldiers exhibit of basic electronics skills. As one example, Soldiers do not fully understand how to apply TMDE in a context that they have not experienced, even if the context is analogous to a situation that they have experienced. The Soldiers seem not to understand the basic mechanisms and functionality of TMDE. The Ordnance school performed a far transfer study involving some 300 94E and 94L Soldiers, where each group had to perform maintenance on a single channel ground and airborne radio system (SINCGARS) radio in a context (aviation or non-aviation) different from what was learned, and using different actual (but not different functional) TMDE, but otherwise involving the same principles. According to several OEMTD instructors, a large number of Soldiers were unable to complete the transfer task.

Methods

Approach

The current work employed categorization as the indirect technique used to derive data for participants' mental models. The justification for using categorization was twofold. First, it was the intent of the research to consider techniques where "[s]ignificant differences between characterizations of any two mental models …[could]… be determined", as between experts and novices. All of the indirect techniques described above fit this criterion, as they all yield data that can be qualitatively and quantitatively analyzed across participants. Second, it was the intent of the research to consider techniques with "ease of use" in an institutional training environment. Categorization perhaps alone among the indirect techniques fits this criterion, as the presentation of stimulus items and capture of categorized items is a straightforward process.

Stimulus Items

The experimental stimulus items used for categorization were taken from several sources, including an OEMTD-provided list of TMDE used by the different MOS at Ft. Gordon and discussions with electronics maintenance experts, primarily instructors and Warrant Officers at Ft. Gordon. As shown in Table 1 on the next page, the research team generated an initial list of 53 possible stimulus items that included three types of items: TMDE identifiers (e.g., *OS-303/G*), TMDE descriptors (e.g., *oscilloscope*, *power meter*), and diagnostic functions (e.g., *voltage measurement*, *power measurement*). Subsequently, the research team, in coordination with OEMTD, finalized the stimulus item list used in the categorization task at 39 items.[1] As a check on the applicability of the stimulus items, and to ensure that items would be relevant to different MOS (since it was expected that different types of electronics maintenance experts, representing different MOS, would take part in the subsequent experiment), a subject-matter expert (SME)[2] who did not take part in the selection of stimulus items was asked to determine the relevance of each item to the different MOS trained at Ft. Gordon OEMTD. Using TMDE technical manuals and MOS descriptions, the SME mapped Soldier tasks to the functions of each piece of test equipment. The SME's mapping is also listed in Table 1.

Participants

As shown in Table 2, there were a total of 83 participants. Novice participants had just begun the basic electronics course and were resident at Ft. Gordon for only a few weeks. Intermediate participants were already part-way through their MOS-specific training and had been resident at Ft. Gordon for an average of about four months. Expert participants were OEMTD instructors and Warrant Officers with years of operational, including deployment, experience. For the latter two groups, the focus was on three of the five MOS: 94E (radio and communications security repairer), 94F (special electronic devices repairer), and 94R (avionic radar repairer). All data gathering occurred at Ft. Gordon.

---

[1] The equipment identifiers were omitted from the experiment as they were deemed unhelpful in mental model derivation, since most novices, and even experts who don't work with those *specific* pieces of equipment, might not know how to categorize the specific identifiers. This was a lesson learned from the effort..

[2] This SME had experience managing and operating TMDE as a Signal NCO and as a Signal Officer, including service as an Airborne Signal Company Commander and service as Chief of Operations in the Airborne Corps G-6.

Table 1
Experimental Stimulus Items

STIMULUS ITEM	MOS RELEVANCY	INCLUDED?
AN/USM-459	94 D	No
AC voltage measurement	94 DEFLR	Yes
ammeter	94 DEFLR	Yes
AN/GRM-122	94 E	No
AN/PSM-45A	94 F	No
AN/URM-213	94 DEF	No
AN/USM-485	94 D	No
AN/USM-486(U)	94 DER	No
AN/USM-488	94 D	No
AN/USM-491	94 R	No
conductance measurement	94 DEFLR	Yes
current measurement	94 DEFLR	Yes
DC voltage measurement	94 DEFLR	Yes
digital multimeter	94 DEFLR	Yes
frequency counter	94 DEFLR	Yes
frequency measurement	94 D	Yes
load match measurement	94 DEL	Yes
microwave frequency counter	94 R	Yes
multimeter	94 DEFLR	Yes
ohmmeter	94 DEFLR	Yes
OS-303/G	94 DEFLR	No
oscilloscope	94 DEFLR	Yes
power measurement	94 DEFLR	Yes
power meter	94 DER	Yes
power to load calculation	94 DER	Yes
pulse generator	94 R	Yes
radar test set	94 R	Yes
radio	94 DE	Yes
radio frequency measurement	94 DER	Yes
radio frequency power test set	94 DER	Yes
radio frequency test set	94 DER	Yes
radio receiver	94 DE	Yes
radio test set	94 DE	Yes
reflectance measurement	94 DE	Yes
resistance measurement	94 DEFLR	Yes
signal amplification	94 DEFLR	Yes
signal distortion	94 DEFLR	Yes
signal generation	94 DEFLR	Yes
signal generator	94 DEFLR	Yes
spectrum analyzer	94 DEFLR	Yes
TD1225A(V)1U	94 R	No
timing measurement	94 DER	Yes
transmission line loss measurement	94 DER	Yes
transmission test set	94 DER	Yes
TS3395A	94 FR	No
TS3662U	94 ELR	No
TS3895A/U(V)	94 FR	No
TS4317/GRM	94 DE	No
voltage measurement	94 DEFLR	Yes
		(Continued)
voltmeter	94 DEFLR	Yes

watt meter	94 DEFLR	Yes
waveform generator	94 FR	Yes
waveform measurement	94 FR	Yes

Table 2
Experimental Participants

GROUP	# OF PARTICIPANTS	EXPERIENCE LEVEL
Novice	21	<1 month
Intermediate	39	3 to 5 months
Expert	23	deployment and/or instructor

Tool

The research staff recommended that the online tool OptimalSort be used for the experiment. This tool allows the participant to drag and drop from a list of stimulus items into either participant-derived (open sort) or experimenter-defined (closed sort) categories. (During data collection, stimulus items are presented randomly to participants; this is an automatic feature of the online tool.) This experiment used an open sort exclusively. Instructions given to all participants are presented in Figure 1. During the experimental sessions, the research staff observed very few difficulties that participants had with using this tool.

Methods

Each group of participants was seated together in a room with each participant assigned to his or her own personal computer. Participants were instructed to read all directions shown on their individual screens, then conduct the categorization task. The tool is designed to be distributed, hence its directions are reasonably comprehensive. Research staff were present during experimental sessions as well to answer questions, though there were few. The categorization task was self-paced, and did not require more than one-half hour for any participant. After all participants completed the categorization task a research staff member debriefed the participants. For the experts, the research staff then led a focused discussion on uses of the categorization approach. Part of this discussion involved having the experts come up with scenarios that could be used in place of or in addition to the categorization task to further clarify a given Soldier's knowledge and skills.

Welcome to OEMTD's online sorting exercise.

Thank you for your participation. This exercise is designed to help us better understand how you think about information relevant to your future job. Your input will help us to provide more effective training.

This survey will take approximately 15 minutes to complete. All results are confidential. Your information will help us to analyze the results of this survey.

To begin, please enter your student ID followed by an @ sign followed by today's date followed by .mil. For example, you might enter FG12345@16SEP08.mil.

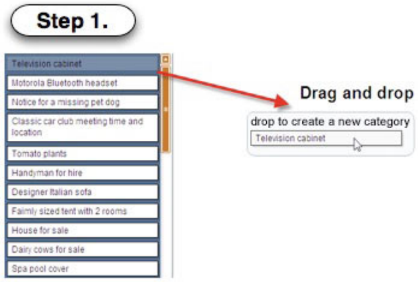
{participant presses the Start button}

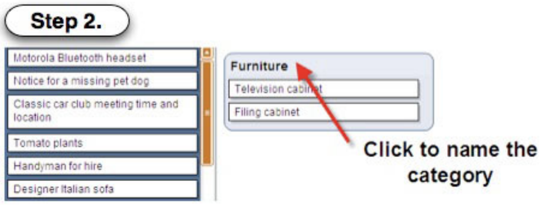
You will see a number of items on the left side of the screen that relate to ordnance test maintenance and diagnostic equipment.

Use your mouse to drag and drop each item into the category you think it belongs to. Drag all the items from the left side into the categories on the right until there are no more items on the left. There is no 'correct' answer, just put the items where you might expect to find them.

To view these instructions at any time during the card sort, click the 'View instructions' link at the top of the page.

- Use your mouse to drag items to your right and form groups that make sense to you. When you're ready, label each group with something you feel comfortable with.
- Drag all the items from the left side into the categories on the right until there are no more items on the left.
- There is no 'correct' answer, just put the items where you might expect to find them.
- To view these instructions at any time during the card sort, click the 'View instructions' link at the top of the page.





To begin, click Continue.

{participant presses the Continue button}

{participant completes the sorting task}

Figure 1. Experimental Instructions Given to All Participants.

RESULTS

Two types of analysis were performed on the categorization data. The first type employed spreadsheets to manipulate the data and derive qualitative and quantitative results. The second type employed statistical methods that were developed by the research team's principal investigator in prior work.

Spreadsheet Analysis

The "Original" column of Table 3 lists the categories generated by the members of the expert group (including categories that were left unlabelled). The same SME who resolved the experimental stimuli (and who did not take part in the conduct of the experiment) was asked to derive a spanning set of labels for the categories that expert participants created in the open sort. The intent here was as a post-experiment check on the stimulus item types. The SME ended up deriving the following labels to cover expert categories: TMDE (specific or unspecified), TMDE function (specific or unspecified), and a general unspecified category ("Standardized Label" column of Table 3). That is, one result – just of expert data – is that, indeed, experts perceived each stimulus item (aside from a few items that were difficult for different experts to categorize hence were thrown into a catch-all category) as belonging to one of two types: TMDE itself (either a specific description of the test equipment or a non-specific description) or TMDE functions (again either a specific description of the function of the test equipment or a non-specific description). This result is unsurprising, since the stimulus items were carefully chosen, but it serves as support for the process taken by the research team of developing a stimulus item list. With the SME's assistance, these labels of expert categories were then mapped to the category names created during the open sort by intermediate and novice participants.

Table 3*SME-derived Standardized Categories*

ORIGINAL	INTERMEDIATE LABEL	STANDARDIZED LABEL
avionics	Components	TMDE - specific
radar	Components	TMDE - specific
radar test equipment	TMDE equipment - testing	TMDE - specific
radar test set equipment	TMDE equipment - testing	TMDE - specific
radio	Components	TMDE - specific
radio equipment	Components	TMDE - specific
radio set equipment	Components	TMDE - specific
advance TMDE	TMDE equipment - testing	TMDE - unspecified
basic TMDE	TMDE equipment - testing	TMDE - unspecified
electronic bench test sets	TMDE equipment - testing	TMDE - unspecified
electronic equip test	TMDE equipment - testing	TMDE - unspecified
equipment	TMDE equipment - testing	TMDE - unspecified
equipment to test	TMDE equipment - testing	TMDE - unspecified
equipment used to measure different electronic variables	TMDE equipment - testing	TMDE - unspecified
Meters	TMDE function - Measurement	TMDE - unspecified
multi-purpose TMDE	TMDE equipment - testing	TMDE - unspecified
not used too often	TMDE equipment - testing	TMDE - unspecified
test equipment	TMDE equipment - testing	TMDE - unspecified

Test Sets	TMDE equipment - testing	TMDE - unspecified
TMDE	TMDE equipment - testing	TMDE - unspecified
types of meters	TMDE equipment - testing	TMDE - unspecified
universal electronic measurement	TMDE equipment - testing	TMDE - unspecified
commo TMDE	TMDE equipment - testing	TMDE function - specific
communication device	TMDE equipment - testing	TMDE function - specific
communications equipment	TMDE equipment - testing	TMDE function - specific
radio tests	TMDE equipment - testing	TMDE function - specific
radio troubleshooting equipment	TMDE equipment - testing	TMDE function - specific
transmission/signal measurement	TMDE function - Measurement	TMDE function - specific
AC/DC measurement	TMDE function - Measurement	TMDE function - specific
ammeter function	TMDE function - Measurement	TMDE function - specific
measure current	TMDE function - Measurement	TMDE function - specific
ohmmeter function	TMDE function - Measurement	TMDE function - specific
voltmeter function	TMDE function - Measurement	TMDE function - specific
counters	TMDE function - spectrum analysis	TMDE function - specific
frequency analysis	TMDE function - spectrum analysis	TMDE function - specific
frequency measurements	TMDE function - spectrum analysis	TMDE function - specific
frequent tests	TMDE function - Measurement	TMDE function - specific
RF counters/measurement equipment	TMDE function - Measurement	TMDE function - specific
signal measurement/analysis equipment	TMDE function - signal analysis	TMDE function - specific
oscilloscope function	TMDE function - signal analysis	TMDE function - specific
power generation	TMDE function - signal analysis	TMDE function - specific
Power/SINAD/impedance matching	TMDE function - Measurement	TMDE function - specific
signal generating device	TMDE function - signal analysis	TMDE function - specific
signal generation equipment	TMDE function - signal analysis	TMDE function - specific
signal generators	TMDE function - signal analysis	TMDE function - specific
signals	TMDE function - signal analysis	TMDE function - specific
signals modifications	TMDE function - signal analysis	TMDE function - specific
RT	TMDE function - spectrum analysis	TMDE function - specific
waveform/signal analysis	TMDE function - spectrum analysis	TMDE function - specific
waves	TMDE function - spectrum analysis	TMDE function - specific
action viewed with TMDE	TMDE function - Measurement	TMDE function - unspecified
analyzer	TMDE function - spectrum analysis	TMDE function - unspecified
automation	Components	TMDE function - unspecified
components to be tested	Components	TMDE function - unspecified
desired data	TMDE function - Measurement	TMDE function - unspecified
electronic measurements that can be adjusted	TMDE function - signal analysis	TMDE function - unspecified
end item specific test set	TMDE equipment - testing	TMDE function - unspecified
generic test set GRM	TMDE equipment - testing	TMDE function - unspecified
measurements	TMDE function - Measurement	TMDE function - unspecified
readings	TMDE function - Measurement	TMDE function - unspecified
the results from testing	TMDE function - Measurement	TMDE function - unspecified
things to be measured	TMDE function - Measurement	TMDE function - unspecified
types of measurements	TMDE function - Measurement	TMDE function - unspecified
what equipment does	TMDE function - Measurement	TMDE function - unspecified
M		unspecified
Misc		unspecified
on system		unspecified
question mark symbol (?)		unspecified
unnamed category 1		unspecified
unnamed category 2		unspecified
unnamed category 3		unspecified

unnamed category 4	unspecified
unnamed category 5	unspecified
unnamed category 6	unspecified
unnamed category 7	unspecified
unnamed category 8	unspecified
unnamed category 9	unspecified
unnamed category 10	unspecified
unnamed category 11	unspecified
unnamed category 12	unspecified
unnamed category 13	unspecified
unnamed category 14	unspecified
unnamed category 15	unspecified
unnamed category 16	unspecified
unnamed category 17	unspecified
unnamed category 18	unspecified
unnamed category 19	unspecified
unnamed category 20	unspecified
unnamed category 21	unspecified
unnamed category 22	unspecified
unnamed category 23	unspecified
unnamed category 24	unspecified
unnamed category 25	unspecified
unnamed category 26	unspecified
unnamed category 27	unspecified
unnamed category 28	unspecified
unnamed category 29	unspecified
unsorted	unspecified

The first analysis of categorization data is a relatively simple one to identify the placement *by experts* of stimulus items into the SME-validated categories. This is a within-group comparison, not yet comparing non-experts against experts. Table 4a shows this data. From the data it can be seen that experts – not unexpectedly – largely agreed with the SME; these indicators are shown as check marks in the last column of the table where agreement achieves 67% , that is, at least two of every three expert participants agreed with the SME. For instance, 77% of the expert participants who categorized *AC voltage measurement* agreed with the SME that it represents a TMDE function, while over 80% of expert participants who categorized *digital multimeter* agreed it represents a description of a kind of TMDE.

By no means, interestingly, does this within-group comparison indicate complete agreement among expert participants, a point that will be brought up again during the discussion of different *types* of experts used for multidimensional scaling analysis. For quite a few stimulus items, a significant percentage of experts categorized the item within some unspecified category. More to point, there are certain items that appear to have bimodal categorization; these items are noted as question marks in the last column, where more than 25% of responses fell into TMDE description categories and at least another 25% fell into TMDE function categories. For instance, for *signal distortion*, most expert participants felt this item represents a TMDE function, however a sizeable percent interpret the item as a description. A similar finding but in the opposite direction holds, for example, for *pulse generator*.

Table 4a

Percent of Expert Groupings that Matched Against SME-Derived Categories

STIMULUS ITEM	SME-DERIVED CATEGORIES			
	UNSPECIFIED	TMDE DESCRIPTION (SPECIFIC OR UNSPECIFIED)	TMDE FUNCTION (SPECIFIC OR UNSPECIFIED)	AGREEMENT BETWEEN SME AND EXPERTS
AC voltage measurement	8%	15%	77%	✓
ammeter	12%	71%	18%	✓
conductance measurement	17%	17%	67%	✓
current measurement	17%	17%	67%	✓
DC voltage measurement	14%	14%	71%	✓
digital multimeter	6%	82%	12%	✓
frequency counter		73%	27%	✓
frequency measurement	15%	15%	69%	✓
load match measurement	15%	8%	77%	✓
microwave frequency counter	6%	69%	25%	✓
multimeter		81%	19%	✓
ohmmeter	12%	71%	18%	✓
oscilloscope		80%	20%	✓
power measurement	14%		86%	✓
power meter	6%	81%	13%	✓
power to load calculation	21%	7%	71%	✓
pulse generator	6%	63%	31%	?
radar test set	13%	80%	7%	✓
radio	13%	38%	50%	?
radio frequency measurement	14%	21%	64%	
radio frequency power test set	20%	53%	27%	?
radio frequency test set	13%	63%	25%	?
radio receiver	13%	38%	50%	?
radio test set	7%	79%	14%	✓
reflectance measurement	21%	7%	71%	✓
resistance measurement	8%	17%	75%	✓
signal amplification	13%	20%	67%	✓
signal distortion	13%	27%	60%	?
signal generation	8%	31%	62%	?
signal generator		71%	29%	✓
spectrum analyzer	6%	75%	19%	✓
timing measurement	15%	8%	77%	✓
transmission line loss measurement	21%	7%	71%	✓
transmission test set	8%	77%	15%	✓
voltage measurement	14%	14%	71%	✓
voltmeter	11%	72%	17%	✓
watt meter	6%	71%	24%	✓
waveform generator		57%	43%	?
waveform measurement	14%	21%	64%	

Overall, the data for 29 of the 39 stimulus items suggest considerable agreement among expert participants, and these items would likely be useful as comparison items for non-expert participant data. The data for another eight of the 39 stimulus items suggest possible disagreement among expert participants, and these data would likely be useful not so much as comparison items for non-expert participant data, but instead as additional information that might be used to dig deeper into a particular non-expert's differences in understanding from experts. This concept is further addressed below. The data for the remaining two stimulus items (*radio frequency measurement* and *waveform measurement*) show neither full nor bimodal agreement among experts; these items are not strongly associated with either TMDE function or description, and would probably be replaced in further tests.

Before comparing non-experts against experts a similar analysis can be done within those groups (i.e., for novice and for intermediate participants), to gauge how high the agreement is among non-experts. The literature would suggest that while experts would tend to agree with each other (supported by these data), intermediate participants, whose mental representations are still developing, would not tend to agree with each other. For instance, Norman et al. (1989, experiment 1) found that patient medical information reading times increased from novice to intermediate participants, then decreased for expert clinicians. They did not, however, find the same results for amount of processed data; amount of data increased across participant groups. Meanwhile, novice participants would focus on superficial aspects of the experimental stimuli and thus perhaps would be expected to show high agreement. For instance, Chi, Feltovich, & Glaser (1981, experiment 1) asked participants to arrange physics problems that included diagrams by similarity of solution procedure. Expert physicists sorted according to physical principle involved, while novices sorted according to irrelevant similarities within diagrams (e.g., objects referred to, or physical configuration). Novices, then, would again tend to agree with each other at the descriptive level but would not understand functional aspects of the TMDE.

Table 4b shows the data for novice participants, with the same criteria for checks and question marks as for experts. It is noted that there were only two bimodal stimulus items (*power to load calculation* and *radio frequency measurement*), and one categorization that is the opposite of the SME's (*signal generation*, noted with an 'x'), results that could be expected when the novices could pay attention only to superficial features of the items. The novices did agree with the SME on 24 of the 39 stimulus items, though, so clearly these novice participants were focused on similar aspects of the stimulus items; a comparison against expert categorizations (done below) will help elucidate *what* it is that the novices were focused on.

Table 4c shows data for the intermediate participants. These data show slightly higher agreement than for expert or novice participants. This finding is somewhat surprising, as some research indicates that intermediate participants have ill-formed mental structures and would thus be likely not to agree with each other. However, it seems reasonable to believe that since all of these participants were midway through intensive electronics maintenance training, they had some good ideas of what TMDE descriptions and functions are, and how to separate those concepts into appropriate categories. To gauge their deeper understanding of TMDE usage, then, the comparison against expert categorization would be telling.

Table 4b

Percent of Novice Groupings that Matched Against SME Derived Categories

STIMULUS ITEM	SME-DERIVED CATEGORIES			
	UNSPECIFIED	TMDE DESCRIPTION (SPECIFIC OR UNSPECIFIED)	TMDE FUNCTION (SPECIFIC OR UNSPECIFIED)	AGREEMENT BETWEEN SME AND NOVICES
AC voltage measurement	14%	10%	76%	✓
ammeter	24%	71%	5%	✓
conductance measurement	14%	14%	71%	✓
current measurement	14%	14%	71%	✓
DC voltage measurement	14%	14%	71%	✓
digital multimeter	19%	62%	19%	
frequency counter	19%	76%	5%	✓
frequency measurement	19%	19%	62%	
load match measurement	14%	14%	71%	✓
microwave frequency counter	24%	76%		✓
multimeter	24%	62%	14%	
ohmmeter	19%	62%	19%	
oscilloscope	19%	67%	14%	✓
power measurement	19%	14%	67%	✓
power meter	19%	62%	19%	
power to load calculation	24%	29%	48%	?
pulse generator	19%	67%	14%	✓
radar test set	14%	86%		✓
radio	19%	76%	5%	✓
radio frequency measurement	14%	29%	57%	?
radio frequency power test set	14%	86%		✓
radio frequency test set	14%	86%		✓
radio receiver	29%	71%		✓
radio test set	14%	86%		✓
reflectance measurement	19%	24%	57%	
resistance measurement	14%	14%	71%	✓
signal amplification	24%	57%	19%	
signal distortion	24%	57%	19%	
signal generation	19%	67%	14%	✗
signal generator	19%	81%		✓
spectrum analyzer	19%	76%	5%	✓
timing measurement	14%	24%	62%	
transmission line loss measurement	19%	19%	62%	
transmission test set	14%	81%	5%	✓
voltage measurement	14%	10%	76%	✓
voltmeter	19%	62%	19%	
watt meter	19%	57%	24%	
waveform generator	29%	67%	5%	✓
waveform measurement	14%	19%	67%	✓

Table 4c
Percent of Intermediate Groupings that Matched Against SME Derived Categories

STIMULUS ITEM	SME-DERIVED CATEGORIES			
	UNSPECIFIED	TMDE DESCRIPTION (SPECIFIC OR UNSPECIFIED)	TMDE FUNCTION (SPECIFIC OR UNSPECIFIED)	AGREEMENT BETWEEN SME AND INTERMEDIATE PARTICIPANTS
AC voltage measurement	14%	11%	75%	✓
ammeter	24%	68%	8%	✓
conductance measurement	21%	7%	73%	✓
current measurement	11%	11%	77%	✓
DC voltage measurement	8%	12%	81%	✓
digital multimeter	8%	85%	8%	✓
frequency counter	13%	79%	9%	✓
frequency measurement	8%	12%	80%	✓
load match measurement	19%	11%	70%	✓
microwave frequency counter	32%	67%	0%	✓
multimeter	4%	88%	8%	✓
ohmmeter	4%	78%	17%	✓
oscilloscope	4%	89%	8%	✓
power measurement	12%	15%	73%	✓
power meter	17%	79%	4%	✓
power to load calculation	28%	24%	48%	
pulse generator	15%	77%	8%	✓
radar test set	28%	65%	7%	
radio	12%	73%	16%	✓
radio frequency measurement	10%	31%	58%	?
radio frequency power test set	12%	76%	12%	✓
radio frequency test set	12%	76%	12%	✓
radio receiver	12%	73%	16%	✓
radio test set	15%	74%	11%	✓
reflectance measurement	20%	10%	70%	✓
resistance measurement	17%	7%	76%	✓
signal amplification	17%	52%	31%	?
signal distortion	17%	45%	38%	?
signal generation	15%	55%	30%	?
signal generator	4%	84%	12%	✓
spectrum analyzer	17%	75%	8%	✓
timing measurement	19%	13%	68%	✓
transmission line loss measurement	22%	13%	66%	
transmission test set	21%	62%	17%	
voltage measurement	11%	15%	74%	✓
voltmeter	5%	82%	14%	✓
watt meter	13%	78%	8%	✓
waveform generator	16%	72%	12%	✓
waveform measurement	4%	15%	81%	✓

Table 5 shows a comparison across all three participant groups for the stimulus items, using markups of Tables 4a,b,c. As this table shows, for fifteen stimulus items all three groups categorized the same, either as a TMDE descriptor or description of its function. Most of these items are basic and well-known TMDE descriptions or functions, such as an *oscilloscope*, a *radio test set*, *power measurement*, and *voltage measurement*. Meanwhile, for eleven additional stimulus items the experts and intermediate participants agree but not the novices. Not surprisingly, these items represent more advanced electronics concepts (*multimeter*, *signal distortion*, *timing measurement*) that intermediate participants could have learned in the several months since they were novices. Another eight stimuli are those that experts differ from the intermediate and novice participants, and likely represent yet more advanced concepts (e.g., *radio frequency test set*, *waveform measurement*) that are not yet well-learned by the time participants reach an intermediate stage. Finally, the remaining five stimulus items yielded either indeterminate categories across the three groups or else no interpretable pattern across the groups; these items seem to be the least informative of the set.

Before turning to statistical analyses, what do these spreadsheet analyses imply for future use with Soldiers? It appears there are certain stimulus items that are understood even at a basic level early on. If any given Soldier does not categorize these items in the same manner as experts then that Soldier may need remediation at that point to learn the basics. Further, there appears to be another subset of items that distinguish intermediate participants from novices, where intermediate participants categorize more similarly to experts and hence may have learned about these TMDE in the time since they were novices. Further still, there appears to be a subset of items that distinguish intermediate from expert participants, suggesting gaps in the knowledge of intermediate participants that would need to be addressed to move them forward towards expertise.

Table 5
Agreement Across Participant Groups

STIMULUS ITEM	EXPERT PARTICIPANTS	INTERMEDIATE PARTICIPANTS	NOVICE PARTICIPANTS
AC voltage measurement	✓	✓	✓
ammeter	✓	✓	✓
conductance measurement	✓	✓	✓
current measurement	✓	✓	✓
DC voltage measurement	✓	✓	✓
digital multimeter	✓	✓	
frequency counter	✓	✓	✓
frequency measurement	✓	✓	
load match measurement	✓	✓	✓
microwave frequency counter	✓	✓	✓
multimeter	✓	✓	
ohmmeter	✓	✓	
oscilloscope	✓	✓	✓
power measurement	✓	✓	✓
power meter	✓	✓	
power to load calculation			
pulse generator		✓	✓
radar test set	✓		✓
radio		✓	✓
radio frequency measurement		✓	✓
radio frequency power test set		✓	✓
radio frequency test set		✓	✓
radio receiver		✓	✓
radio test set	✓	✓	✓
reflectance measurement	✓	✓	
resistance measurement	✓	✓	✓
signal amplification			
signal distortion	✓	✓	
signal generation	✓	✓	
signal generator	✓	✓	✓
spectrum analyzer	✓	✓	✓
timing measurement	✓	✓	
transmission line loss measurement			
transmission test set	✓		✓
voltage measurement	✓	✓	✓
voltmeter	✓	✓	
watt meter	✓	✓	
waveform generator		✓	✓
waveform measurement		✓	✓

Statistical Analysis

Participants' categorization data were subjected to a multidimensional scaling (MDS) analysis, described next. This analysis largely supported the previous findings that, for the most part, participants all used somewhat similar criteria for sorting, but there are quantitative means for identifying certain telling differences.

MDS provides a view of how participants perceive experimental stimuli; that is, what underlying features they perceive as important. MDS analyses demand that the analyst specify certain parameters, for instance, a reasonable value for number of *dimensions*. The procedure will return a "stress" or "badness-of-fit" value, which should be minimized, a value somewhat analogous to variance unaccounted for in an analysis of variance. Generally, with more dimensions this badness-of-fit lessens; however, because dimensions are assumed to correspond to psychological dimensions, the difficulty of labeling dimensions increases with the number of dimensions (see Schiffman, Reynolds, & Young, 1981). Similarly, MDS allows the experimenter to specify whether or not individual participants can weight dimensions differently. When they can, as was allowed here, the experimenter may analyze the weights themselves in an attempt to find regularities.

To run any MDS analysis, first, all pairwise comparisons for all stimuli for all participants are calculated, simply with a value of 1 for commonly-grouped stimulus items and 0 for pairs of stimulus items not grouped together. The resultant table is sometimes called a similarity matrix. Thus, to be specific, for the two stimulus items *ammeter* and *ohmmeter*, if a given participant dragged and dropped them together under one category (of the participant's choosing), then for that participant's similarity matrix that pairing was assigned a 1, while if instead the participant placed them into different categories (of his/her choosing) then that pairing was assigned a 0. From these data can be run multiple MDS analyses, varying dimensionality, to assess a reasonable (and interpretable) number of underlying dimensions to the data. For these data successive runs were made using two, three, four, and five underlying dimensions to the categorization data. Dimensions higher than three did not yield results appreciably more informative than three dimensions, so the analysis continued with three dimensional weightings.

Qualitative ordering of stimulus items. A simple first analysis of these data is qualitative and involves just sorting the stimuli based on a particular dimension's weights, then looking for patterns in the arrangement of the stimuli. The art to an MDS analysis is to determine what dimensions imply. Not all dimensions will portray any obvious pattern; just because they are analytically feasible doesn't mean they are psychologically interpretable. To conduct this artful exercise the same SME, who was not otherwise involved in any of the statistical data analyses, was asked to determine if there were any patterns for the three dimensional weightings of the experts.

Table 6 lists the stimulus items arranged according the three dimensional weights, that is, the columns are arrayed according to the value each item takes on each dimension that comes out of the MDS, suggesting what criteria experts used to categorize the items. It is apparent from Column A that, for the most part, descriptions of TMDE cluster towards one end of the list and functions of TMDE cluster towards the other end. The SME discerned a different pattern in the

arrangement of items in Column B, conjecturing that the items showed an ordering geared towards how the various test equipment is used. For Column C, meanwhile, the data appeared to the SME to be grouped using the logic of evaluation. That is, when receiving a faulty piece of equipment a Soldier can never take for granted what is wrong with it, so as a technician s/he needs to evaluate the equipment to determine the problem(s), and only once that determination is made go on to diagnose the problem to determine the correct course of action to take. To the SME, the arrangement of items in this third column reflected the iterative actions of use of a piece of test equipment and obtaining results that inform subsequent activities.

Table 6
Stimulus Items Sorted According to Experts' Dimensions

COLUMN A. STIMULUS ITEMS (& DIMENSION WEIGHTS) SORTED ACCORDING TO EXPERTS' FIRST DIMENSION.	COLUMN B. STIMULUS ITEMS (& DIMENSION WEIGHTS) SORTED ACCORDING TO EXPERTS' SECOND DIMENSION.	COLUMN C. STIMULUS ITEMS (& DIMENSION WEIGHTS) SORTED ACCORDING TO EXPERTS' THIRD DIMENSION.
microwave freq. counter (-1.81)	radio (-2.16)	waveform generator (-3.4)
frequency counter (-1.45)	radar test set (-1.94)	signal distortion (-1.72)
pulse generator (-1.45)	power to load calc. (-1.78)	resistance measurement (-1.46)
digital multimeter (-1.37)	radio freq. power test set (-1.39)	signal generator (-1.37)
radio test set (-1.2)	radio receiver (-1.36)	signal generation (-1.3)
radio frequency test set (-1.07)	signal amplification (-1.24)	signal amplification (-1.05)
ohmmeter (-1.05)	pulse generator (-1.22)	pulse generator (-0.97)
signal generator (-0.99)	transmission test set (-0.96)	waveform measurement (-0.63)
radio freq. power test set (-0.98)	radio freq. measurement (-0.76)	frequency counter (-0.58)
radar test set (-0.95)	signal distortion (-0.76)	timing measurement (-0.56)
spectrum analyzer (-0.89)	radio frequency test set (-0.7)	oscilloscope (-0.51)
multimeter (-0.83)	radio test set (-0.62)	spectrum analyzer (-0.43)
voltmeter (-0.68)	transm. line loss meas. (-0.31)	frequency measurement (-0.37)
oscilloscope (-0.66)	signal generation (-0.25)	radar test set (-0.3)
power meter (-0.53)	reflectance measurement (-0.23)	power to load calc. (-0.29)
watt meter (-0.34)	timing measurement (-0.22)	transm. line loss meas. (-0.18)
ammeter (-0.22)	spectrum analyzer (-0.14)	voltage measurement (-0.14)
signal generation (-0.16)	voltage measurement (-0.04)	microwave freq. counter (-0.05)
transmission test set (-0.15)	watt meter (-0.04)	watt meter (0.14)
waveform generator (-0.08)	power measurement (0.01)	radio (0.15)
transm. line loss meas. (0.14)	load match measurement (0.13)	power meter (0.18)
radio (0.17)	voltmeter (0.18)	AC voltage measurement (0.26)
radio receiver (0.19)	waveform measurement (0.18)	digital multimeter (0.35)
signal amplification (0.27)	signal generator (0.32)	load match measurement (0.43)
reflectance measurement (0.33)	microwave freq. counter (0.33)	radio freq. power test set (0.49)
signal distortion (0.33)	resistance measurement (0.37)	radio test set (0.54)
waveform measurement (0.41)	frequency counter (0.77)	conductance meas. (0.62)
DC voltage measurement (0.83)	ohmmeter (0.82)	multimeter (0.65)
current measurement (0.9)	conductance meas. (0.86)	radio freq. measurement (0.7)
frequency measurement (0.96)	current measurement (0.95)	DC voltage measurement (0.71)
AC voltage measurement (1.05)	digital multimeter (0.97)	radio receiver (0.82)
voltage measurement (1.08)	frequency measurement (1.17)	reflectance measurement (0.9)
conductance meas. (1.24)	ammeter (1.26)	ammeter (1.01)
power to load calc. (1.31)	waveform generator (1.27)	power measurement (1.01)
radio freq. measurement (1.34)	DC voltage measurement (1.28)	ohmmeter (1.02)
resistance measurement (1.49)	AC voltage measurement (1.29)	current measurement (1.06)
power measurement (1.51)	multimeter (1.29)	radio frequency test set (1.16)
load match measurement (1.62)	oscilloscope (1.33)	transmission test set (1.22)
timing measurement (1.7)	power meter (1.33)	voltmeter (1.88)

Quantitative comparison across groups. A more complicated analysis of MDS results is quantitative. For this form of analysis, the dimension weights themselves for each participant or participant group are subjected to an analysis of variance (ANOVA), and systematic differences between groups are sought. This analysis is similar to lining up experts', intermediate participants', and novices' weightings against each other and looking for any correlation, except that the stimulus items as well could be additionally labeled according to different attributes (e.g., as either descriptive or functional types), and any differences among groups for those attributes sought.

The model used for the ANOVA assessed dimensional weightings according to participant group, stimulus item type, and their possible interaction. For the first dimension, this model accounted for $R^2=0.75$ ($SS_{Model}=88.19$, $SS_{Corrected\ Total}=117.06$) of the total variance. According to the ANOVA, the group weightings for the first dimension (i.e., as shown in Column A of Table 6, that used by experts to separate descriptive from functional stimulus items) showed no effect across participant groups ($F(2,111)<1$, $MSE=0.0$) but large effects for stimulus type (descriptive or functional) alone ($F(1,111)=32.97$, $p<0.01$, $MSE=8.58$) and for the interaction between participant group and stimulus type ($F(2,111)=153.03$, $p<0.01$, $MSE=39.81$). The significant finding for stimulus type alone simply supports the qualitative analysis; in essence the dimensional weights for descriptive items group towards one end and the dimensional weights for functional items group towards the other end. Put another way, along this first dimension, descriptions of TMDE tended toward one direction while descriptions of functions of TMDE tended toward the other direction. This finding simply implies that one of the important underlying aspects of the stimuli that all participants noticed was the distinction between description and function of TMDE, supporting the spreadsheet analyses described above. Meanwhile, the significant finding for the interaction between participant group and stimulus type was further investigated through planned contrasts. These analyses showed that experts differed from novices for functional items but not descriptive items, that experts differed from intermediate participants for descriptive items but not functional items, and that intermediate participants differed from novice participants for both descriptive and functional items. One explanation behind these findings is uncomplicated: Novices are able to identify TMDE by description but not by function whereas experts are able to organize by both description and function, while intermediate participants are still forming their mental models of TMDE.

Meanwhile, for the second and third dimensions the model accounted for none of the total variance (both $R^2<1$), and there were no effects across participant groups nor for stimulus type nor for the interaction between participant group and stimulus type (all $F\leq 1.23$, *ns.*), all suggesting that these dimensions reflect different reasons (i.e., not according to either descriptive or functional stimulus types) for the stimulus item arrangement, as was hinted by the SME's interpretation of how the items were arrayed in the latter two columns of Table 6.

Refined quantitative comparison for specific participants. A more instructive quantitative analysis, however, is not to compare *all* non-experts against *all* experts, but instead compare *a given* non-expert against different groups or types of experts. This approach, after all, moves closer to the goal of being able to model the structure and process of knowledge and skills exhibited by a particular Soldier and to identify differences between his/her model and one or more experts' models. Hence the SME was asked to identify types of experts based on how they developed groupings (primarily by considering their category labels). Because the experiment

was run using an open (rather than closed) sort, participants were able to create their own categories. For the experts this capability was illuminating, as it enabled the SME to better understand the different approaches experts could take in categorizing the stimuli.

To make these different approaches clear, the SME carefully analyzed how the expert participants in this experiment grouped stimuli. The SME considered both what items were grouped together and what labels, if any, the expert participants applied. Out of this analysis the SME derived three main approaches to categorization (see Table 7 for examples). The first type of experts might be called “diagnosticians”, as their groupings and labels suggested that they perceived the experimental stimuli as reflecting either different types of diagnostic test equipment or the outcomes from using test equipment. The second type of experts might be called “appliers”, as their groupings and labels suggested that they categorized the stimuli based on how they would differently use the test equipment, such as for radio communications versus radar applications. The third type of experts might be called “functional”, as their groupings and labels suggested a categorization based on the stimuli being either the test equipment itself or descriptive of the function of the equipment. The three types of experts (diagnosticians, appliers, and functional) are seen as aligning, respectively, with the structural, behavioral, and functional representational views of Hmelo-Silver and Pfeffer (2004) presented earlier.[3]

Table 7
SME-derived Expert Groups

GROUP	REPRESENTATIVE CATEGORY LABELS
Diagnostician expert #1	the results from testing test equipment equipment to test what equipment does
Diagnostician expert #2	things to be measured Equipment used to measure different electronic variables electronic measurements that can be adjusted Electronic Bench Test sets Components to be Tested
Appliers expert #1	Universal Electronic Measurement Frequency Analysis Waveform/Signal Analysis

[3] An examination was done to see if any MOS associated with any of the expert types. For the experts whom the SME labeled ‘diagnosticians’, one 94F and one 94R were represented, but since only two experts fell into this group no implications can reasonably be drawn. However, for the experts whom the SME labeled ‘appliers’, 94D, E, & F, and a 948B Warrant, were all represented, suggesting that experts regardless of type (i.e., specialty implied by the experts’ MOS) can tend to view test equipment in terms of how it is employed. Furthermore, for the experts whom the SME labeled ‘functional’, only 94E (five of them) and 94R (two of them) were represented, suggesting that certain experts, due to the specialized nature of their work (e.g., working mainly with communications or radar equipment) may tend to focus on the functions of the equipment needed for their work.

	Power Generation
	Transmission/Signal Measurement
	End Item Specific Test Set
Appliers expert #2	Desired Data
	Basic TMDE
	Advance TMDE
	Action viewed with TMDE
	Commo TMDE
Functional expert #1	Readings
	Meters
	Equipment
	signal generators
	Test Sets
Functional expert #2	Measurements
	TMDE
	Multi-purpose TMDE
	Communications Equipment
	Signals

From this new view of experts being of three types the categorization data were reanalyzed. The same modeling approach as above was used for the ANOVA. For the first dimension, this model accounted for R^2=0.73 (SS_{Model}=85.17, $SS_{Corrected\ Total}$=117.04) of the total variance. On the first dimension, there were no statistical differences among these three expert types alone ($F(2,111)<1$, MSE=0.0) but there *were* differences for both stimulus item type alone ($F(1,111)=39.46$, $p<0.01$, MSE=11.33) and for the interaction between expert type and stimulus type ($F(2,111)=128.56$, $p<0.01$, MSE=36.92). As before and as would be expected, all tests for the second and third dimensions yielded non-significant results (both $R^2<1$ and all $F<1$). For the first dimension, the planned contrasts for the interaction make more clear how the different types of experts categorized different types of stimuli. For instance, diagnosticians and functional experts differed in how they categorized functional stimulus items (that is, those items defining the functions of TMDE) ($F(1,111)=92.48$, $p<0.01$, MSE=26.56) but not in how they categorized descriptive stimulus items (that is, those items identifying TMDE) ($F(1,111)<1$, MSE<1). Conversely, diagnosticians and appliers differed not in how they categorized functional stimulus items ($F(1,111)<1$, MSE<1) but in how they categorized descriptive stimulus items ($F(1,111)=124.64$, $p<0.01$, MSE=35.79), suggesting that the two groups viewed the items similarly, in how they would use the TMDE.

This same statistical comparison can be made for any group – or individual. Of particular interest is then testing any given *non*-expert against the different types of experts to more comprehensively identify how the non-expert differs from different experts. By running an ANOVA on that non-expert's dimensional weights versus the expert types' weights, significant differences and planned contrasts become telling, indicating how exactly the individual differs from experts. Thus, for example, one random intermediate Soldier's data was compared against the different experts, using a model similar to that above that considered type of participant (non-expert vs. three types of experts), stimulus item type, and their interaction. The results indicate that along the first dimension, the model explains R^2=0.69 (SS_{Model}=107.72, $SS_{Corrected\ Total}$=156.07) of the variance, and this particular participant's categorization of functional and descriptive items differs ($F(1,148)=146.80$, $p<0.01$, MSE= 47.95) from the

functional experts but not from the diagnosticians ($F(1,148)=2.33$, *ns.*, MSE<1) or appliers ($F(1,148)<1$, MSE<1). Similarly, one random novice Soldier's data was compared against the different experts, and the results indicate that along the first dimension, the same model explains $R^2=0.55$ ($SS_{Model}=85.25$, $SS_{Corrected\ Total}=156.04$) of the variance, and this particular participant's categorization of functional and descriptive items differed (all $F(1,148)\geq 10.16$, $p<0.01$, all MSE≥4.86) from all three types of experts. Such findings could then inform tailored instruction for that Soldier.

Discussion

The statistical procedures described in this report are sufficient to differentiate the categorization performed by one participant (e.g., a non-expert Soldier) against that of another participant or group of participants (e.g., several expert instructors who categorize similarly). The process is straightforward, entailing: (1) having the Soldier perform a categorization task, possibly using an online tool requiring no more than one-half hour depending on the number and complexity of stimuli; (2) converting the Soldier's resultant groupings into a similarity matrix by assigning values of 1 or 0 to stimulus item pairs that do or do not end up in the same group; (3) generating dimensional weightings using MDS analysis; (4) and running an ANOVA to compare the Soldier's dimensional weightings against the comparison group's. Planned contrasts would then inform the observer/instructor where specific differences in categorization lie.

Spreadsheets with formulas to produce the similarity matrix are available as a product of this research effort.[4] Additionally, programs to generate MDS dimensional weightings and run analyses of variances are available as a product of this effort.[5]

Two decisions to be made that require some foresight involve the stimulus items to use for categorization and the number of MDS dimensions to generate. The stimuli used in this research were designed with one attribute – being a description of TMDE or a description of function of TMDE – in mind but other attributes (e.g., usage of TMDE, as evidenced by the SME's analysis of the arrangement of stimuli in Column B of Table 6 according to experts' dimensional weightings) might suggest adding or replacing stimulus items. As a rule two or three dimensions will suffice, depending on how many attributes each stimulus item takes, as beyond three dimensions it might be difficult to interpret the different dimensions.

Lastly, a decision must be made to use either closed sort or open sort. Given the exploratory nature of this research, an open sort demanding the labeling of categories enabled the researchers to capture additional information from participants. However, for 'production' runs a closed sort using pre-established categories would streamline the processes of assessing a participant's mental model.

---

[4] A spreadsheet model was used because the online tool generated a spreadsheet of categorization data, and because it was a convenient approach to producing the qualitative analyses necessitated by the use of an open sort. A program can easily be written to convert closed-sort categorization data directly into a similarity matrix but this would require access to the raw data from the participant's categorization, which was not available to the research staff.

[5] Statistical programs are written for SAS.

Reliability, Validity, And Usability Of The Mental Models And Mental Modeling Technique

Reliability in this context refers to the ability of a given technique to produce the same model – having only "minimal" differences – under similar circumstances. The analytic approach presented here appears to attain sufficient reliability in that different experts, for instance those of the diagnostic type, were able to produce similar category groupings that resulted in similar dimensional weighting. Further, qualitative spreadsheet analyses support many of the findings of MDS analyses. Further still, analysis of experts' data either altogether or separately by expert types yielded similar dimensions, including one dimension distinguishing between the description of TMDE and the description of functions of TMDE and another dimension arraying stimulus items as associated with their usage. A categorization task followed by MDS and analysis of dimensional weightings offers promise as a reliable technique specifically because the process can be used to enable different participants to demonstrate similar groupings (hence similar mental models of the function of stimulus items) or the same participant to demonstrate noticeable differences in his or her groupings (hence a revised mental model) that result from learning.

Validity in this context refers to the ability of a given technique to discriminate between experts and novices, and, when there are different types of experts (Bransford, Brown, & Cocking, 1999; Murphy & Wright, 1984), to discriminate between experts. Again, the analytic approach presented here appears to attain sufficient validity in that there are different types of expert with qualitatively and quantitatively different dimensional weightings. Further, non-expert weightings can be compared against the different experts' to find regularities in the differences. Additionally, some of the findings in this research mirror those found elsewhere using non-categorization tasks (e.g., Norman et al., 1989), such as studies to demonstrate how intermediate participants' mental models are developing and appear neither like experts' nor like novices'.

Usability in this context refers to the ability of training providers to integrate mental modeling into their training environments. Though beyond the scope of this effort, the Ordnance School is looking towards training Soldiers not only in TMDE procedures but also TMDE concepts. There may be a resulting change in curricula, or specific course modules may need to be developed. Given a reliable, valid means for the instructors to characterize Soldiers' mental models of TMDE functionality before, during, and after training (relative to experts' mental models), instructors and instructional developers might be able to tailor TMDE training.

Additional Research

The research team offers three areas that might represent useful and interesting future research. First, in this research only textual materials, that is words or short phrases, were used as stimulus items, a limit imposed by the online tool that may be removed in future versions of the tool or by use of another categorization tool. Diagrams in particular, but other stimulus forms (e.g., images of TMDE or of TMDE outputs) too, have been used in past research (e.g., Chi et al., 1981; Vicente, 1992) to better understand participants' mental models particularly of complex equipment such as is appropriate to Ordnance maintenance technicians.

Second, and related to the use of non-textual stimulus items, is use of mini scenarios describing validation-like exercises that might serve as additional mental modeling materials. In other work by the research team (e.g., Hubal et al., 2008) scenarios are used to situate the participant in a context appropriate for assessing target skills. Similarly, a slightly more complex but ecologically valid task could have Soldiers producing the same categorized data to feed an MDS analysis.

Third, a kind of sensitivity analysis of stimulus items might prove valuable. Not only would it be of value for practitioners to understand how many stimuli are necessary but also which are sufficient. This analysis could also help instructors and instructional developers understand the types of domains that have well-defined stimuli that would be appropriate for this methodology. The attributes assigned to stimuli, as shown above with the description/function attribute, can inform the ANOVA of dimensional weightings and make clear specific differences between non-experts and experts (or among any other individuals or groups). Any given set of stimuli will of course be domain-specific, but a sensitivity analysis using stimuli such as natural categories (Rosch, 1973) where much is already known about the stimuli can inform decisions of number and characteristics of stimulus items.

References

Barnard, Y., Reiss, M., & Mazoyer, P. (2006). Mental models of aircraft maintenance documentation. In F. Reuzeau, K. Korker, & G. Boy (Eds.), *Proceedings of the International Conference on Human-Computer Interaction in Aeronautics* (pp. 232-239). Toulouse, France: Cépaduès-Editions.

Chase, W.G., & Simon, H.A. (1973). Perception in chess. *Cognitive Psychology*, *4*, 55-81.

Chi, M.T.H., Feltovich, P.J., & Glaser, R. (1981). Categorization and representation of physics problems by experts and novices. *Cognitive Science*, *5*, 121-152.

Chiesi, H.L., Spilich, G.J., & Voss, J.F. (1979). Acquisition of domain-related information in relation to high and low domain knowledge. *Journal of Verbal Learning and Verbal Behavior*, *18*, 257-273.

Committee on Developments in the Science of Learning (1999). How experts differ from novices. In J.D. Bransford, A.L. Brown, & R.R. Cocking (Eds.), *How people learn: Brain, mind, experience, and school* (pp. 17-38). Washington, DC: National Academies Press.

Cooke, N.J., & Schvaneveldt, R.W. (1988). Effects of computer programming experience on network representations of abstract programming concepts. *International Journal of Man-Machine Studies*, *29*, 407-427.

Frank, G., Hubal, R., & O'Bea, M. (2007). Using competency definitions to adapt training for mission success. In *Proceedings of the Interservice/Industry Training, Simulation and Education Conference* (pp. 1262-1270). Arlington, VA: National Defense Industrial Association.

Frank, G.A., Helms, R.F., & Voor, D. (2000). Determining the right mix of live, virtual, and constructive training. In *Proceedings of the Interservice/Industry Training, Simulation and Education Conference* (pp. 1268-1277). Arlington, VA: National Defense Industrial Association.

Freyhof, H., Gruber, H., & Ziegler, A. (1992). Expertise and hierarchical knowledge representation in chess. *Psychological Research, 54*, 32-37.

Goel, A., Mahesh, K., Peterson, J., & Eiselt, K. (1996). Unification of language understanding, device comprehension and knowledge acquisition. In *Proceedings of the Annual Conference of the Cognitive Science Society* (pp. 540-545), July 12-15, 1996, La Jolla, CA.

Gott, S.P., Bennett, W., & Gillet, A. (1986). Models of technical competence for intelligent tutoring systems. *Journal of Computer-Based Instruction*, *13*, 43-46.

Hagemann, K., & Scholderer, J. (2007). Consumer versus expert hazard identification: A mental models study of mutation-bred rice. *Journal of Risk Research*, *10*, 449-464.

Hanisch, K.A., Kramer, A.F., & Hulin, C.L. (1991). Cognitive representations, control, and understanding of complex systems: A field study focusing on components of users' mental models and expert/novice differences. *Ergonomics*, *34*, 1129-1145.

Hassebrock, F., Johnson, P.E., Bullemer, P., Fox, P.W., & Moller, J.H. (1993). When less is more: Representation and selective memory in expert problem solving. *American Journal of Psychology*, *106*, 155-189.

Hmelo-Silver, C.E., & Pfeffer, M.G. (2004). Comparing expert and novice understanding of a complex system from the perspective of structures, behaviors, and functions. *Cognitive Science*, *28*, 127-138.

Hubal, R.C., Fishbein, D.H., Sheppard, M.S, Paschall, M.J., Eldreth, D.L., & Hyde, C.T. (2008). How do varied populations interact with embodied conversational agents? Findings from inner-city adolescents and prisoners. *Computers in Human Behavior*, *24*, 1104-1138.

Kieras, D.E., & Bovair, S. (1984). The role of a mental model in learning to operate a device. *Cognitive Science*, *8*, 255-273.

McGraw, K.M., & Pinney, N. (1990). The effects of general and domain-specific expertise on political memory and judgment. *Social Cognition*, 8, 9-30.

Murphy, G.L., & Wright, J.C. (1984). Changes in conceptual structure with expertise: Differences between real-world experts and novices. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *10*, 144-155.

Norman, D.A. (1983). Some observations on mental models. In D. Gentner & A. Stevens (Eds.), *Mental Models* (pp. 7-14). Hillsdale NJ: Erlbaum.

Norman, D.A. (1988). *The psychology of everyday things*. New York, NY: Basic Books.

Norman, G.R., Brooks, L.R., & Allen, S.W. (1989). Recall by expert medical practitioners and novices as a record of processing attention. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *15*, 1166-1174.

Rosch, E.H. (1973). Natural categories. *Cognitive Psychology*, *4*, 328-350.

Schiffman, S.S., Reynolds, M.L., & Young, F.W. (1981). *Introduction to multidimensional scaling*. New York: Academic Press.

Smith, M.U. (1990). Knowledge structures and the nature of expertise in classical genetics. *Cognition and Instruction*, *7*, 287-302.

Vicente, K.J. (1992). Memory recall in a process control system: A measure of expertise and display effectiveness. *Memory & Cognition*, *20*, 356-373.

Weiser, M., & Shertz, J. (1983). Programming problem representation in novice and expert programmers. *International Journal of Man-Machine Studies*, *19*, 391-398.

APPENDIX A

LIST OF ACRONYMS

ANOVA	Analysis of variance
ARI	U.S. Army Research Institute for the Behavioral and Social Sciences
CAI	Computer aided Instruction
G-6	Chief Information Officer
MDS	Multidimensional scaling
MOS	Military occupational specialty
NCO	Noncommissioned officer
OEMTD	U.S. Army Ordnance Electronics Maintenance Training Directorate
RTI	RTI International
SINCGARS	Single channel ground and airborne radio system
SME	Subject-matter expert
TMDE	Test, measurement, and diagnostic equipment